Report of Data Mining Homework 3, Anomaly Detection

109550027, 紀竺均

1. Explain your implementation which gets the best performance in detail.
    a. Data Preprocessing
    I use pandas to load the dataset and use StandardScaler to normalize the features columns
    b. Build the Autoencoder
    The autoencoder has three main parts: an encoder, a bottleneck, and a decoder. The encoder compresses the input, the bottleneck represents the compressed knowledge, and the decoder reconstructs the input from the bottleneck.
    Activation Function: I have tried different activation functions for intermediate layers such as Relu, LeakyReLU, and ELU, finally, I choose ReLU for its lowest validation loss.
    Loss Function: I use MSE to measure the reconstruction error.
    Optimizer: Choose Adam optimizer because of its effectiveness in various conditions.
    Autoencoder Summary

Model: "functional_17"

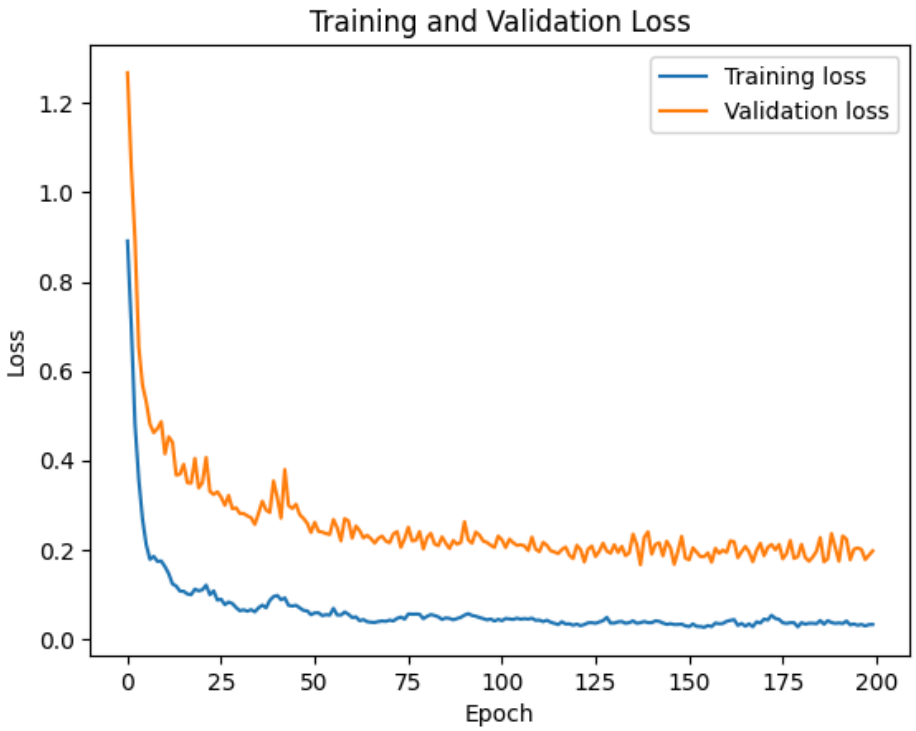
Layer (type)	Output Shape	Param #
input_layer_8 (InputLayer)	(None, 16)	0
dense_48 (Dense)	(None, 128)	2,176
dense_49 (Dense)	(None, 64)	8,256
dense_50 (Dense)	(None, 32)	2,080
dense_51 (Dense)	(None, 64)	2,112
dense_52 (Dense)	(None, 128)	8,320
dense_53 (Dense)	(None, 16)	2,064

Total params: 25,008 (97.69 KB)

Trainable params: 25,008 (97.69 KB)

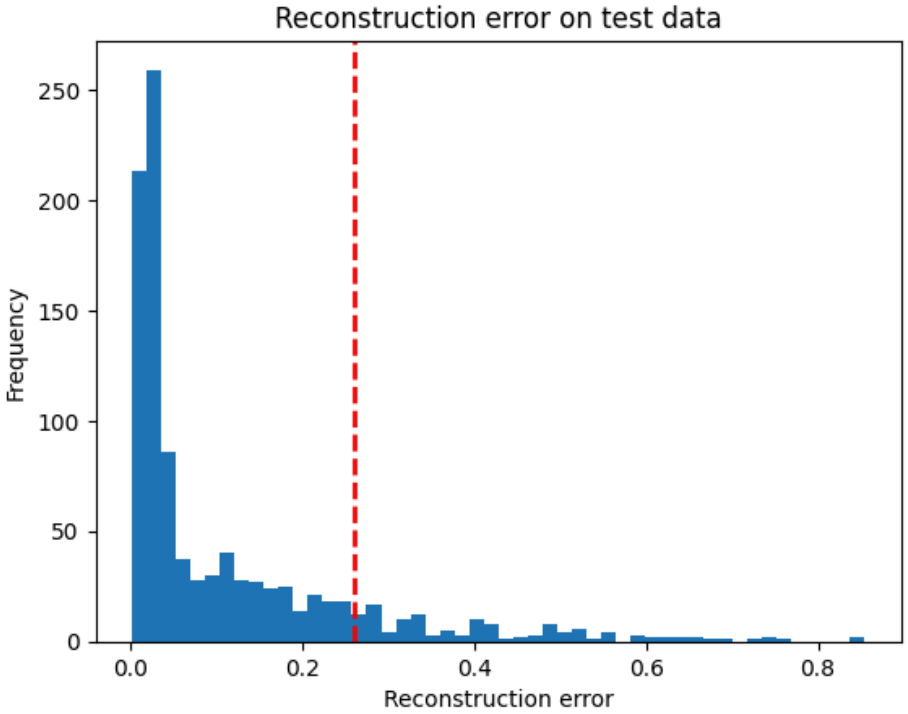
Non-trainable params: 0 (0.00 B)

    c. Monitor Training Process
    Use autoencoder.fit() to train the model, monitoring the training loss and validation loss during the process.



d. Evaluation

Compute the reconstruction error (MSE) on the training and testing data. Set the threshold of the training_MSE to 95% and detect the anomalies if testing_mse is higher than the threshold.



2. Explain the rationale for using AUC score instead of F1 score for binary classification in this homework.
    a. **Imbalance Classes**: AUC is more robust in imbalance datasets, where one class is more prevalent than another. In our homework, the anomalies(outliers) are rarer (400:600) than the normal one, so AUC is preferred.
    b. **Performance Across All Thresholds**: AUC evaluates how well the model can distinguish between the two classes, independent of how the decision threshold is set. In this assignment, the T.A.s want to evaluate the model's ability to discriminate between classes across all possible thresholds, using AUC gives a more comprehensive view of model performance.

3. Discuss the difference between semi-supervised learning and unsupervised learning.

    Unsupervised learning is used when no labels are available. It focuses on identifying patterns and structures in purely input data, while **semi-supervised learning** uses a small amount of labeled data along with a large amount of unlabeled data during training to build more accurate models. The presence of labeled data provides a basis for evaluating and guiding the learning process; it is particularly useful when labels are limited or expensive to produce.